Chapitre 1 Représentation des données complexes pour la fouille

Résumé. Au cours de ce chapitre nous posons la problématique de la représentation d'un document sous une forme exploitable par les méthodes usuelles de fouille de données. Nous montrons comment l'extraction de caractéristiques au sein de documents complexes permet cette mise en forme. Nous discutons de la difficulté du choix du grain informationnel et des descripteurs. Nous décrivons des descripteurs usuels pour la représentation de données temporelles, images, vidéos et textuelles. Nous rappelons ensuite le principe général des méthodes sélectionnant les descripteurs et celles permettant la construction de descripteurs synthétiques.

Nous avons étudié plus particulièrement le cas des données textuelles pour lesquelles nous discutons de divers pré-traitements permettant une meilleure exploitation et de l'apport des traitements linguistiques. Nous discuterons également des méthodes de sélection de termes, problématique pour laquelle nous avons proposé une nouvelle méthode.

Mots clefs : Extraction de caractéristiques, sélection de variables, données images, données vidéos, données textuelles.

Chapitre 1 – Représentation des données complexes

1.1 Introduction

La fouille de données, ou *data mining* en anglais, est l'analyse des observations de larges jeux de données dans le but d'identifier des relations non soupçonnées et de résumer la connaissance incluse au sein de ces données sous de nouvelles formes à la fois compréhensibles et utiles pour l'expert de ces données (FAYYAD *et al.*, 1996). Selon KODRATOFF (1994) et rappelée par ZIGHED *et al.* (2000, p. 28), la fouille de données n'est qu'une étape d'un processus plus global appelé Extraction de Connaissances à partir de Données (ECD), ou *Knowledge Discovery from Databases* (KDD) en anglais : l'ECD se réfère à une démarche complète d'exploitation des données intégrant leur pré-traitement pour permettre l'application des algorithmes de fouilles de données suivie de la validation des modèles obtenus parvenant ainsi au stade de connaissances.

Comme nous l'avons déjà décrit, le pré-traitement des données est une phase cruciale puisque du choix des descripteurs et de la connaissance précise de la population va dépendre la mise au point des modèles de prédiction. L'information nécessaire à la construction d'un bon modèle de prévision peut être disponible dans les données mais un choix inapproprié de variables ou d'échantillons d'apprentissage peut faire échouer l'opération (ZIGHED et RAKOTOMALALA, 2002).

Cette étape de mise en forme dépend fortement de la nature du document traité et a pour principe d'extraire des caractéristiques, descripteurs ou encore variables exogènes. Cette représentation est généralement nécessaire pour pouvoir utiliser les méthodes de fouilles de données opérant sur des tables bidimensionnelles.

Dans le cadre des données complexes, cette opération est encore plus délicate. Un objet complexe peut être considéré comme un agrégat de documents complexes. Ces relations peuvent contenir elles-mêmes de l'information. En reprenant l'exemple du dossier médical, une quantité de clichés radiographiques fortement supérieure à la moyenne associée à un faible taux de calcium peut indiquer que ce patient a des difficultés à fixer le calcium sur ses os et nécessite un traitement particulier. Les systèmes de gestion de bases de données (SGBD) et les fichiers XML sont des moyens permettant de gérer ce niveau d'information, considéré dans ce cas comme de la méta-information. Les SGBD le font à travers la définition d'un modèle relationel mettant en évidence les différentes interactions entre les éléments. Les fichiers XML le font à partir d'une grammaire définissant ainsi la composition des différents éléments en sous-éléments, eux-mêmes décomposés pour arriver à des données élémentaires.

En outre, un objet complexe peut être composé de documents de nature différente. Typiquement, nous retrouvons ce genre de situation sur le Web : nous pouvons considérer comme objet complexe un site web. Ce dernier est composé d'un ensemble de pages web dont l'architecture peut être informative. De plus, chacune de ces pages peut elle-même comporter du texte, des images, du son, … L'étape de représentation doit alors décrire le contenu informationnel intrinsèque à chaque document et selon sa nature sous une forme vectorielle.

Dans le cadre d'un corpus d'objets complexes, la représentation doit alors être cohérente sur l'ensemble du corpus. En effet, il faut qu'elle soit capable de représenter de façon identique deux objets identiques. De même, deux objets du corpus fortement différents devront avoir des représentations fortement dissemblables au sens d'une mesure de similarité à définir.

Après avoir poser nos définitions et notations dans la section suivante (1.2) nous décrirons en section 1.3 les caractéristiques usuellement utilisées dans le cadre de documents complexes comme les images ou les textes. Nous ne prétendons pas à une présentation exhaustive des différentes caractéristiques car d'une part il en existe pléthore pour chacun des types de documents abordés et d'autre part car nous ne sommes pas experts dans leur traitement propre (e.g. traitement d'images). Nous recherchons ici à faire émerger les motivations et la démarche dans la recherche de ces caractéristiques pour extraire de la connaissance. C'est pourquoi nous aborderons ensuite le fonctionnement général des méthodes de sélection et construction de variables dans les sections 1.4 et 1.5.

Enfin, nous nous intéresserons en section 1.6 à un document complexe particulier : les données textuelles, pour lesquelles nous appronfondierons la recherche de caractéristiques. Nous décrirons dans un premier temps quelles sont les spécificités de telles données afin de comprendre le contenu informationnel qu'elles recèlent. Dans un second temps nous détaillerons des méthodes de sélection de variables textuelles qui comme nous le verrons se trouve être une étape particulièrement nécessaire. A cette occasion, nous présenterons une nouvelle méthode de sélection que nous comparerons aux autres.

1.2 Définitions et notations

Une population Ω composée de n individus ou objets est décrite par un ensemble de descripteurs $\Gamma = \left\{X^1,\ldots,X^p\right\}$, appelé dans le domaine de la statistique ensemble de variables exogènes. Ces varia-

bles prennent leurs valeurs dans un espace de représentation noté $\Re$ ne possédant pas de structure mathématique particulière :

$$\begin{aligned} \Gamma : \Omega &\mapsto \Re \\ i &\to \Gamma(i) = \left(x_i^1, \ldots, x_i^p\right) \end{aligned} \tag{1.2.1}$$

L'ensemble de ces valeurs est recueilli pour chacun des individus de Ω au sein d'une matrice de données notée *D*. Elle est donc de dimensions $n \times p$.

Dans un cadre d'apprentissage supervisé, où nous cherchons à établir un diagnostic ou établir des prévisions, une variable particulière est associée à la population Ω. Elle est notée E et est nommée indifféremment classe, endogène ou simplement étiquette. Elle prend ses valeurs dans l'ensemble des étiquettes noté $\mathcal{E} = \{e_1, \ldots, e_m\}$:

$$\begin{aligned} \mathrm{E} : \Omega &\mapsto \mathcal{E} = (e_1, \ldots, e_m) \\ i &\to \mathrm{E}(i) \end{aligned} \tag{1.2.2}$$

Deux cas se distinguent selon que les étiquettes sont exclusives ou non :

- Dans le premier cas, un individu est supposé n'appartenir qu'à une et une seule étiquette. Il s'agit alors de prédire à laquelle de ces étiquettes appartient l'individu, ce qui revient à travailler avec une seule variable endogène catégorielle, notée Y et prenant ses valeurs dans $\mathcal{E}$. Ainsi, $\mathrm{E} = Y$;
- Dans le second cas, un individu peut appartenir à aucune, une ou plusieurs étiquettes. Il s'agit alors de prédire à quelles classes appartient le document, ce qui revient à travailler avec *m* variables notées $Y^1, \ldots, Y^m$; d'où $\mathrm{E} = \left(Y^1, \ldots, Y^m\right)$. La variable Y^k prend la valeur $y_i^k = 1$ pour l'individu *i* si il appartient à l'étiquette e_k , 0 sinon. Plus généralement, $y_i^k \in [0,1]$ et définit alors le degré d'appartenance (e.g. une probabilité) de l'individu *i* à l'étiquette e_k .

L'objectif de la fouille de données est alors d'élaborer un modèle φ , appelé encore classifieur, à partir des variables exogènes permettant de prédire à quelle(s) étiquette(s) est associé un individu. Cet aspect sera détaillé dans le Chapitre 2. Plus formellement, le classifieur est donc une application de $\Re$ dans l'ensemble des étiquettes $\mathcal{E}$:

$$\begin{aligned} &\varphi : \Re \mapsto \mathcal{E} \\ &\Gamma(i) \to \varphi(\Gamma(i)) \end{aligned} \tag{1.2.3}$$

Remarque. La variable exogène $j \in \Gamma$ est identifiée au vecteur X^j à *n* composantes. Pour alléger les notations, nous substituerons *j* à X^j lorsqu'il n'y aura pas d'ambiguïté. La variable endogène $k \in \mathrm{E}$ est identifiée au vecteur Y^k à *n* composantes. Enfin, lorsqu'il n'y aura qu'une seule variable endogène et pas d'ambiguïté, nous substituerons *k* à l'étiquette $e_k \in \mathcal{E}$.

1.3 Définition des caractéristiques

La définition ou extraction de caractéristiques au sein d'un document complexe est comme nous l'avons vu une étape cruciale puisque la représentation qui en découle doit conserver au mieux l'information contenue dans le document. Une remarque naïve serait de dire qu'il suffit d'utiliser les éléments primitifs de chacun des types de données pour conserver l'information. Pour les données temporelles il s'agirait d'un évènement au sein d'une séquence, pour les images des valeurs de pixels et pour les textes des lettres.

Ces éléments constituent ce que l'on peut appeler des grains informationnels composant un document. En prenant l'exemple des données textuelles, le plus petit grain informationnel se trouve être le caractère. A un niveau supérieur nous rencontrons le mot, englobant un ensemble de caractères. Puis, à un niveau plus global, nous pouvons définir les phrases, les paragraphes, … et pour finir le document lui-même. LEBART et SALEM (1994) appellent ces grains, dans le cadre du texte, des unités sémantiques.

La difficulté est donc le choix du grain puisque ce dernier influe directement sur le niveau d'information que l'on cherche à recueillir. Prenons cette fois-ci un exemple sur les données images. D'un point de vue de la perception, la couleur d'un pixel n'est que faiblement informative. En revanche, dire qu'une zone de pixels composée d'un pixel de couleur blanche encadré de pixels de couleur noire peut signifier quelque chose.

Il nous semble qu'une réponse générale quant au choix du grain informationnel serait insatisfaisante tant les problèmes à traiter n'utilisent pas le même niveau d'information. Néanmoins, deux pistes sont actuellement exploitées pour pallier ces problèmes. La première aborde la question d'un point de vue sémantique en se basant sur des éléments informatifs à l'humain comme les syntagmes nominaux ou les paragraphes pour le cas des textes ou des objets contenus dans les images. La seconde aborde la

question d'un point de vue statistique en se basant sur l'extraction de fragments contenant de l'information à faire émerger telle l'utilisation des n-grammes dans les textes et des matrices de co-occurrences pour les images.

A partir de ces éléments informationnels, nous calculons des caractéristiques. Ce sont ces dernières que nous allons décrire dans les paragraphes suivants.

1.3.1 Données temporelles

Différentes caractéristiques utiles peuvent être extraites des données temporelles. La modélisation markovienne suppose qu'un événement *X* survenu à l'instant *t* peut être prédit en fonction des *k* évènements précédents $\{X_{t-1}, X_{t-2}, \cdots, X_{t-k}\}$, encore appelés retards. Dès lors, nous pouvons représenter un individu comme étant composé de *k* variables exogènes, les retards, et d'une variable endogène, l'événement à l'instant *t*. Le système ainsi caractérisé est dit auto-régressif.

Si les données se trouvent sous forme de séquences indépendantes et de longueur variable, nous pouvons construire des primitives permettant d'extraire pour chacune de ces séquences les *k* retards de l'événement à l'instant *t*. Pour ce faire, il suffit seulement d'extraire ces éléments selon un fenêtrage de taille *k+1* et de se déplacer d'un élément (cf. Figure 1-1).

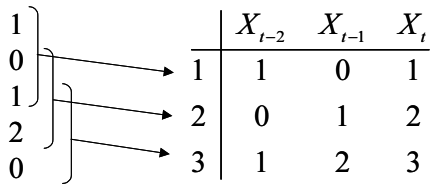

Figure 1-1 Exemple de mise en forme d'une séquence

1.3.2 Données images

Pour extraire les caractéristiques d'une image *I*, nous l'interprétons comme une matrice de dimensions $N \times M$, où *N* est le nombre de lignes et *M* le nombre de colonnes, et dont les éléments $I_{x,y}$ correspondent aux valeurs des pixels de l'image (cf. Figure 1-2).

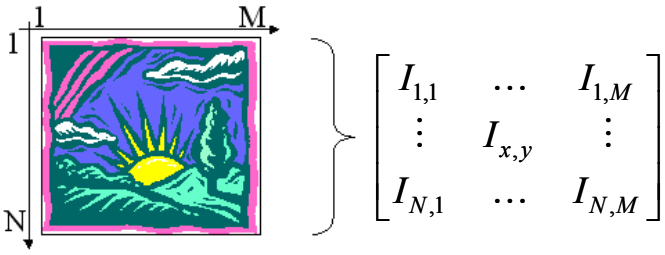

Figure 1-2 Lecture matricielle d'une image

Dans le cadre des données images, le choix des caractéristiques à calculer peut s'effectuer à deux niveaux :

- Le premier concerne la globalité de l'image, auquel nous pouvons extraire des caractéristiques de couleur (ou niveau de gris) modélisant le spectre visuel, et des caractéristiques de texture représentant les détails de surface des objets réels (SCUTURICI, 2002) ;
- Le second concerne les objets au sein de l'image (cf. Figure 1-3), auxquels nous pouvons extraire les mêmes caractéristiques que précédemment, mais en outre nous pouvons extraire des caractéristiques rendant compte des spécificités de la forme et du contour comme par exemple le périmètre, l'aire, la circularité, le degré d'élongation (LONCARNIC, 1998). Néanmoins, l'extraction d'objets est une étape délicate faisant appel aux techniques de segmentation d'images. Cette décomposition de l'image en objets est l'objectif de la norme MPEG-7 (TELECOMITALIALAB), mais à ce jour, il n'existe pas de logiciels d'extraction automatique d'objets mettant en œuvre cette norme. Par ailleurs, cette décomposition en objets nous conduit à redéfinir la représentation matricielle du corpus. En effet, le nombre d'objets d'une image à une autre est vraisemblablement variable et distinct. Nous traitons ce problème en disant que l'individu que nous allons utiliser est un objet d'une image, un objet pouvant potentiellement être l'image elle-même. Nous retrouvons une représentation fortement similaire à *D*. Cependant ici *n* ne correspond plus au nombre de documents du corpus mais au nombre d'objets. Par ailleurs, l'ajout d'une variable index permet de déterminer à quelle image appartient un objet. Nous ne ferons donc plus la distinction entre des caractéristiques globales ou liées à un objet d'une image.

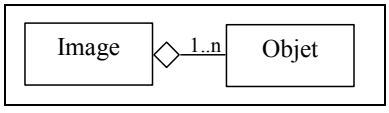

Figure 1-3 Décomposition élémentaire d'une image

Nous donnons ci-dessous les quelques caractéristiques que nous avons utilisées lors de nos expérimentations.

1.3.2.1 Caractéristiques de couleur

Comme l'illustre la Figure 1-4, dans le cas des images couleurs, à chaque pixel $I_{x,y}$ de I est associé le triplet $I_{x,y} = \left(I^{r}{}_{x,y}, I^{v}{}_{x,y}, I^{b}{}_{x,y} \right)$ correspondant aux trois canaux de couleur utilisés. Ici, nous nous sommes basés sur le système RVB (respectivement Rouge, Vert et Bleu), mais d'autres modèles existent (FOLEY *et al.*, 1995, pp. 584-598). Nous ferons remarquer que dans le cas d'une image en niveaux de gris, nous lui associerons la même valeur pour chacun des trois canaux, ceci afin de ne pas différencier le traitement des images couleur de celles à niveaux de gris.

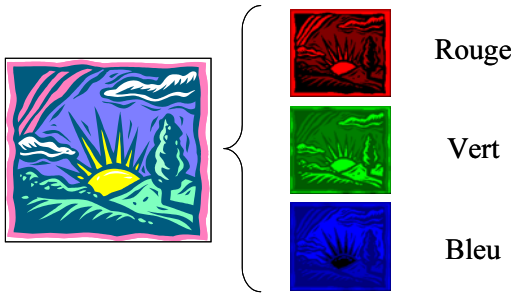

Figure 1-4 Décomposition d'une image en 3 canaux : Rouge, Vert et Bleu

Par ailleurs, à chacun de ces canaux correspond un ensemble de niveaux dans lequel évolue la valeur du pixel. Le nombre de niveaux sera noté $n^{\max}$ et considéré identique pour les trois canaux.

Les formules des caractéristiques de couleurs que nous allons présenter sont applicables sur chacun des canaux utilisés, donc nous ne parlerons pas d'un canal en particulier mais bien de chacun d'entre eux. Par ailleurs, les caractéristiques utilisées dans nos travaux ont été décrites dans (SCUTURICI, 2002) :

Couleur prédominante. Soit p_k le nombre de fois où le niveau de gris k apparaît dans l'image I : $p_k = \left| \left\{ \forall (x,y) \in N \times M, I_{x,y} = k \right\} \right|$, où la notation $|A|$ désigne le cardinal d'un ensemble A. Alors, la couleur prédominante correspondant à la plus grande valeur absolue des valeurs de pixels est $C = \underset{k \in [0, n^{\max}-1]}{\operatorname{argmax}} \left(p_k \right)$;

Norme L1, Norme L1 normalisée. La norme L1 d'une image correspond à la somme des valeurs de pixels : $\|I\|_{L1} = \sum_{x,y \in N \times M} I_{x,y}$; et sa normalisation se fait en divisant la norme L1 par le nombre de pixels de l'image : $\|I\|_{\widetilde{L1}} = \frac{1}{N.M} \sum_{x,y \in N \times M} I_{x,y}$;

Norme L2, Norme L2 normalisée. La norme L2 d'une image correspond au radical de la somme des carrés des valeurs des pixels : $\|I\|_{L2} = \sqrt{\sum_{x,y \in N \times M} I_{x,y}^{\ 2}}$. De même que précédemment, la normalisation s'effectue en divisant la norme par le nombre de pixels de l'image : $\|I\|_{\widetilde{L2}} = \frac{1}{N.M} \sqrt{\sum_{x,y \in N \times M} I_{x,y}^{\ 2}}$;

1.3.2.2 Caractéristiques de texture

La texture est l'une des plus importantes caractéristiques utilisées pour identifier des objets ou des régions d'intérêt dans une image. Elle contient des informations sur la distribution spatiale ou statistique des couleurs. Les caractéristiques de texture ont la propriété de discriminer différentes régions. En se basant sur l'hypothèse que l'information de texture d'une image donnée I est représentée par l'organisation spatiale des niveaux de couleurs de I les uns par rapport aux autres. HARALICK *et al.* (1973) proposent le calcul de 14 paramètres de texture à partir des matrices de co-occurrences.

Matrice de co-occurrences. Une matrice de co-occurrences mesure la probabilité d'apparition des paires de valeurs de pixels situés à une certaine distance dans l'image. Elle est basée sur le calcul de la probabilité $P(i,j,\delta,\theta)$ suivant : $P(i,j,\delta,\theta)$ représente le nombre de fois où un pixel de niveau de couleur i apparaît à une distance relative δ d'un pixel de niveau de couleur j et selon une orientation θ donnée.

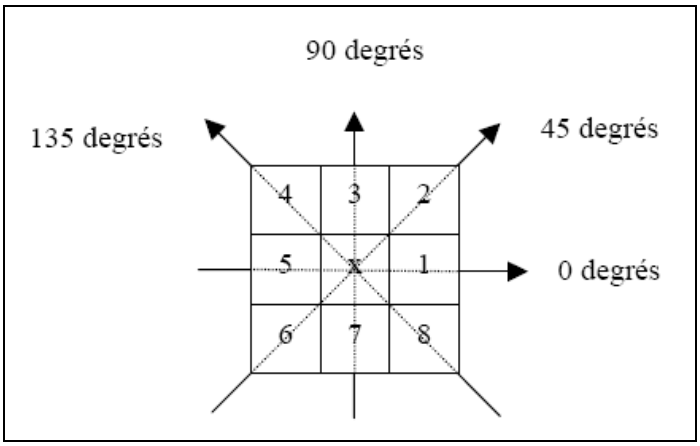

Figure 1-5 Plus proches voisins du pixel 'x' selon 4 directions

Les directions angulaires θ classiquement utilisées sont 0, 45, 90 et 135 degrés. Les relations de voisinage entre pixels, nécessaires au calcul des matrices, sont illustrées en Figure 1-5 ; par exemple, les plus proches voisins de 'x' selon la direction $\theta = 135$ degrés sont les pixels 4 et 8.

Les caractéristiques extraites à partir de ces matrices contiennent des informations notamment sur l'homogénéité, les dépendances linéaires entre les niveaux de gris, le contraste et la complexité de cette image.

Les matrices obtenues selon ces quatre directions sont alors calculées comme dans (1.3.1), (1.3.2), (1.3.3) et (1.3.4) où (k,l) sont les coordonnées d'un pixel de niveau de couleur $i \in \left[0, n^{\max} - 1\right]$ et (m,n) celles du pixel de niveau de couleur $j \in \left[0, n^{\max} - 1\right]$:

$$P(i,j,\delta,0) = \left| \left\{ \left((k,l),(m,n)\right) \in (N \times M)^2 \setminus \left(k - m = 0, |l - n| = \delta, I_{k,l} = i, I_{m,n} = j\right) \right\} \right| \tag{1.3.1}$$

$$P(i,j,\delta,45) = \left| \left\{ \begin{array}{l} \left((k,l),(m,n)\right) \in (N \times M)^2 \setminus \\ \left(k - m = \delta, l - n = -\delta\right) \vee \left(k - m = -\delta, l - n = \delta\right), I_{k,l} = i, I_{m,n} = j \end{array} \right\} \right| \tag{1.3.2}$$

$$P(i,j,\delta,90) = \left| \left\{ \left((k,l),(m,n)\right) \in (N \times M)^2 \setminus \left(|k - m| = \delta, l - n = 0, I_{k,l} = i, I_{m,n} = j\right) \right\} \right| \tag{1.3.3}$$

$$P(i,j,\delta,135) = \left| \left\{ \begin{array}{l} \left((k,l),(m,n)\right) \in (N \times M)^2 \setminus \\ \left(k - m = \delta, l - n = \delta\right) \vee \left(k - m = -\delta, l - n = -\delta\right), I_{k,l} = i, I_{m,n} = j \end{array} \right\} \right| \tag{1.3.4}$$

Ces matrices sont ensuite normalisées par le nombre total de paires R d'une matrice donnée, d'où pour une direction θ et une distance δ données : $\tilde{P}(i,j) = \dfrac{P(i,j)}{R}$.

Pour calculer les paramètres de HARALICK *et al.* (1973), définis en Annexe A, nous utilisons les notations suivantes :

Soient μ_x, μ_y, σ_x et σ_y les moyennes et écart-types des distributions marginales $\tilde{P}_x = \sum_{j=1}^{n^{\max}} \tilde{P}(i,j)$ et $\tilde{P}_y = \sum_{i=1}^{n^{\max}} \tilde{P}(i,j)$ associées à $\tilde{P}(i,j)$.

Nous considérons aussi :

$$\tilde{P}_{x+y}(k)=\sum_{i=1}^{n^{\max}}\sum_{\substack{j=1\\ i+j=k}}^{n^{\max}}\tilde{P}(i,j) \text{ pour } k=2,3,\ldots,2n^{\max} \text{ et } \tilde{P}_{x-y}(k)=\sum_{i=1}^{n^{\max}}\sum_{\substack{j=1\\ |i-j|=k}}^{n^{\max}}\tilde{P}(i,j) \text{ pour } k=0,1,\ldots,n^{\max}-1.$$

Certaines de ces caractéristiques ont des propriétés facilement compréhensibles :

- Le paramètre f_1 mesure l'homogénéité de l'image ;
- f_2 mesure les variations locales des niveaux de couleurs présentes dans une image ou la région étudiée, telles que plus il y a de variations, plus la valeur du contraste f_2 sera élevée ;
- L'entropie f_9 représente le caractère aléatoire des valeurs des niveaux de couleurs. Lors d'une grande variabilité entre les éléments de la matrice, l'entropie est faible alors qu'elle est sensiblement élevée quand les éléments sont égaux.

1.3.2.3 Caractéristiques statistiques

Les *moments de l'image* sont des caractéristiques statistiques. Ils sont calculés à partir des valeurs des pixels ainsi que de leur position relative dans l'image. Ils peuvent être considérés comme des caractéristiques de couleur car utilisant les valeurs des pixels et non des paires de pixels comme dans le cas de la texture. Pourtant, les moments peuvent être aussi utilisés comme des caractéristiques sur la forme lorsqu'ils sont calculés sur le contour ou la surface d'une région d'intérêt de l'image (TUCERYAN, 1994) :

Moment spatial, moment spatial normalisé. Soient x,y les coordonnées du pixel $I_{x,y}$ et $\gamma,\eta\in\mathbb{N}$ tel que $0\le\gamma+\eta\le 3$ alors $\gamma+\eta$ est appelé l'ordre du moment de l'image I, et le moment spatial est : $M_U(\gamma,\eta)=\sum_{x\in[1,N]}\sum_{y\in[1,M]}x^{\gamma}y^{\eta}I_{x,y}$ et le moment spatial normalisé est : $\bar{M}_U(\gamma,\eta)=\frac{M_U(\gamma,\eta)}{N^{\gamma}M^{\eta}}$;

Moment central, moment central normalisé. Soient $\bar{x}=\frac{M_U(1,0)}{M_U(0,0)}$ et $\bar{y}=\frac{M_U(0,1)}{M_U(0,0)}$ les coordonnées du centre de gravité, alors le moment central de I est : $U_U(\gamma,\eta)=\sum_{x\in[1,N]}\sum_{y\in[1,M]}(x-\bar{x})^{\gamma}(y-\bar{y})^{\eta}I_{x,y}$ et le moment central normalisé est : $\bar{U}_U(\gamma,\eta)=\frac{U_U(\gamma,\eta)}{N^{\gamma}M^{\eta}}$.

En outre, des moments particuliers ont une signification géométrique :

- $M_U(0,0) = U_U(0,0)$ représente le périmètre ou la surface de la région considérée ;
- $M_U(0,0)$, $M_U(1,1)$ et $M_U(2,0)$ peuvent être utilisés notamment pour déterminer l'élongation d'une région et ses axes principaux.

1.3.3 Données vidéos

Les données vidéos sont des documents multimédia par essence. Elles sont caractérisées par la présence de contenus visuels (e.g. séquence d'images), sonores (e.g. bande-son, dialogue) et textuels (e.g. sous-titrage) ; un document vidéo est donc en lui-même un document complexe.

La modélisation d'un document vidéo contient un aspect temporel et un aspect structurel :

- L'aspect temporel est du à son principe même d'élaboration ; les données vidéos sont composées d'une succession d'images fixes. Représenter des données vidéos en considérant cet aspect revient à extraire des caractéristiques d'une séquence (cf. 1.3.1) composée d'évènements. Ces évènements étant les images fixes, les caractéristiques les composant (cf. 1.3.2) peuvent être utilisées. Cependant, cet aspect temporel est à prendre en considération seulement si la problématique l'impose ;
- D'un point de vue structurel, la vidéo peut être considérée comme composée d'une hiérarchie de sous-objets comme l'illustre la Figure 1-6 (SCUTURICI, 2002). Chaque document vidéo est composé d'une ou plusieurs scènes, qui sont composées d'un ou plusieurs plans, qui contiennent plusieurs images et chaque image contient un ou plusieurs objets.

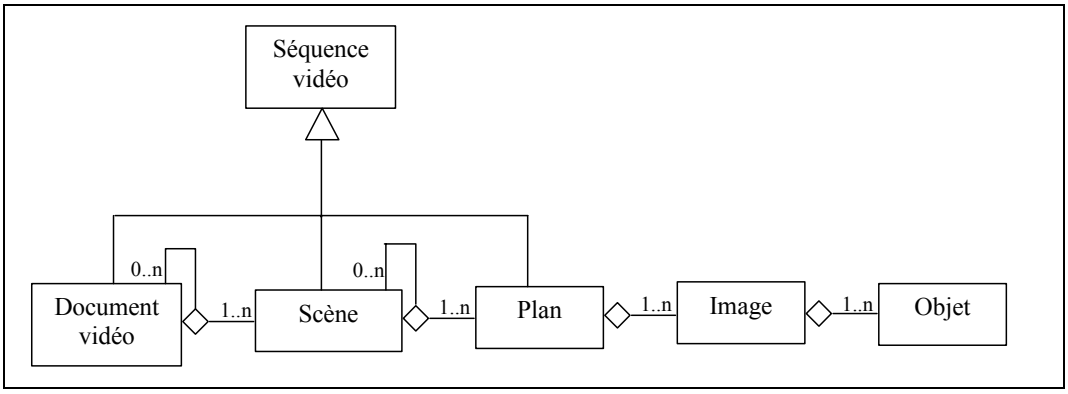

Figure 1-6 Les unités structurelles d'un document vidéo

Plan. Un plan est une série continue d'images qui représente une action continue dans le temps et l'espace. Les images qui composent un plan sont donc enregistrées sans interruption de l'enregistreur (e.g. caméra).

Scène. Une scène vidéo est un ensemble d'images ou de plans représentant un concept sémantique. Une définition similaire est donnée par (MAHDI, 2001) : « Une scène est un ensemble de plans ayant une unité narrative qui peut être définie selon le lieu ou l'action et qui décrit généralement le même sujet ou partage les mêmes objets ».

La difficulté dans la représentation de données vidéo est de déterminer l'individu à extraire au sein de la vidéo. Travailler directement avec l'ensemble des images composant la séquence semble d'un faible intérêt puisque d'une part le nombre d'images serait très élevé (e.g. de l'ordre de 25 images par seconde) et qu'une majorité d'entre elles serait fortement ressemblante puisque deux images successives d'un même plan sont faiblement différentes.

Cette dernière remarque pourrait dans certains cas motiver à considérer comme individu plutôt le plan que l'image. Pour ce faire, il faut être capable de détecter le début et la fin de chaque plan composant la vidéo. A ce titre, plusieurs méthodes existent et détectent la majorité des changement de plans. Une méthode se basant sur la remarque précédente indique la faible variation de deux images consécutives d'un même plan. La contraposée de cette hypothèse revient à considérer que la variation entre deux images consécutives appartenant à deux plans différents est élevée. Cette variation peut se mesurer à travers la différence de la somme des pixels composant les images. Enfin, il faut choisir un ensemble restreint d'images représentatives de la séquence. Nous pouvons par exemple considérer à cet effet la première, celle du milieu et la dernière image composant le plan (ZHANG *et al.*, 1995), effectuer des traitements plus élaborés et déterminer l'image la plus caractéristique du plan, ou encore de s'appuyer sur une mosaïque d'images permettant ainsi une représentation plus compacte du contenu du plan (IRANI et ANANDAN, 1998).

L'utilisation des scènes permet d'incorporer un niveau sémantique plus important. Elles peuvent intervenir en tant qu'étiquette d'un problème de catégorisation, où la problématique serait d'identifier automatiquement à quelle scène appartient un certain plan.

1.3.4 Données textuelles

LUHN (1958) dans ses travaux sur le résumé automatique posa l'hypothèse que la fréquence des occurrences d'un mot au sein d'un texte fournissait une mesure pratique de l'importance de ce dernier. Par ailleurs, un document est couramment synthétisé par un index (VAN RIJSBERGEN, 1979). Plusieurs

travaux ont été mis en œuvre afin de déterminer le choix le plus approprié pour élaborer ces index allant de l'utilisation délicate d'un langage relationnel (FARRADANE *et al.*, 1973) à la simple extraction de termes. L'interrogation sous-jacente à ces travaux était l'efficacité de l'exploitation des index constitués automatiquement face à leur élaboration manuelle. Des études menées, comme (CLEVERDON *et al.*, 1966 ; AITCHISON *et al.*, 1970), ont montré que les index basés sur un contrôle du vocabulaire n'obtenaient pas plus de meilleurs résultats que ceux constitués automatiquement. A cette vision, SALTON (1968) ajouta la pondération des termes formant les index afin d'ajuster l'information qu'ils contiennent en fonction par exemple de la taille du document.

Dans la suite de cette section, nous allons discuter du choix du descripteur (1.3.4.1) puis des différentes pondérations possibles (1.3.4.2).

1.3.4.1 Nature des descripteurs

Différentes méthodes sont proposées pour le choix des termes et les poids associés à ces termes (YANG, 1999). SALTON et MCGILL (1983) et AAS et EIKVIL (1999) utilisent les mots comme descripteurs, d'autres préfèrent l'utilisation de groupes de mots comme les mots composés, les expressions, ou encore les syntagmes nominaux (SIDHOM, 2002). Ces descripteurs ont pour avantage évident de posséder un sens explicite. Cependant, plusieurs problèmes se posent.

En effet, il faut tout d'abord définir ce qu'est un mot pour pouvoir le traiter automatiquement. On peut le considérer comme étant une suite de caractères appartenant à un dictionnaire, ou de façon plus pragmatique comme étant une séquence de caractères non-délimiteurs encadrés par des caractères délimiteurs, couramment les caractères de ponctuation (GILLI, 1988), ce qui renvoie au problème de la gestion des sigles et mots composés pouvant nécessiter un pré-traitement linguistique.

Un autre problème dans le choix des mots en tant que descripteurs est la présence de mots outils et de mots porteurs de sens. Les premiers, constitués des articles, déterminants, …, constituent une grande part des mots d'un texte, mais sont faiblement informatifs pour déterminer le contenu du document. Usuellement, ces mots sont recueillis dans des ante-lexiques pour chacune des langues étudiées et sont supprimés systématiquement du corpus. Malheureusement, il se peut qu'un élément soit considéré comme mot outil dans certains cas et comme porteur de sens dans d'autres.

Enfin, la différenciation effective des mots d'une même famille en raison de leur variation morphologique (e.g. déménageur, déménageurs, déménagement, déménagements, déménager, déménage) est

généralement un inconvénient car engendrant des fréquences très faibles pour chacun de ces termes, alors que les regrouper permettrait d'avoir un nombre d'occurrences plus important et d'amoindrir le phénomène d'imprécision des fréquences rappelée dans (JALAM, 2003). Les procédures de lemmatisation et de stemming tentent de résoudre ces problèmes :

- La lemmatisation cherche à obtenir la racine linguistique du mot pour remplacer par exemple les verbes par leur forme infinitive et les noms par leur forme au singulier. Cette méthode fait appel à l'étiquetage grammatical comme celui de BRILL (1994). Un algorithme efficace incluant un étiqueteur grammatical, nommé *Tree Tagger* (SCHMID, 1994), a été développé pour les langues anglaise, française, allemande et italienne. La principale spécificité de cet outil est que l'étiqueteur repose sur un arbre de décision modélisé par un corpus d'apprentissage dans chacune des langues citées ;
- L'extraction des stems repose quant à elle sur des contraintes linguistiques bien moins fortes en se basant essentiellement sur la morphologie flexionnelle mais aussi dérivationnelle. De ce fait, les algorithmes sont beaucoup plus simplistes et mécaniques que ceux permettant l'extraction des lemmes, engendrant deux avantages : plus rapides que la lemmatisation (il n'y a pas d'étiquetage à effectuer) et la capacité à traiter les mots inconnus sans traitement spécifique. Plusieurs algorithmes ont été proposés ; l'un des plus connus pour la langue anglaise est l'algorithme de PORTER (1980), actuellement décliné pour traiter une vingtaine de langues dont le français. Il existe d'autres algorithmes que celui de PORTER pour déterminer les racines lexicales : une comparaison entre différents algorithmes a été menée dans (HULL, 1996).

L'utilisation du stemming peut apporter des résultats supérieurs à ceux obtenus par lemmatisation comme montrés dans (DE LOUPY, 2001) dans un cadre de recherche documentaire. Mais il convient de signaler que cette opération peut mener à des confusions sémantiques en regroupant accidentellement deux mots sémantiquement différents, et que ces deux procédures (stemming et lemmatisation) sont dépendantes de la langue, nécessitant de les adapter pour chaque langue utilisée.

Il existe une autre approche de la représentation des textes : les n-grammes (SHANNON, 1948). Un n-gramme est une séquence de *n* caractères (CAVNAR et TRENKLE, 1994). Il est à noter qu'il existe un autre sens où cette fois-ci, un n-gramme est une suite de *n* mots et dont l'utilisation principale est de capturer les relations locales entre les mots en déterminant quel mot va apparaître conditionnellement à la présence des n-1 mots précédents (BROWN *et al.*, 1992). Dans la totalité de ce document, nous utilisons n-gramme uniquement dans le sens de séquence de *n* caractères.

Pour un document quelconque, l'ensemble des n-grammes pouvant être généré est obtenu en déplaçant une fenêtre de *n* caractères sur l'ensemble du texte. Ce déplacement s'effectue caractère par caractère. A chaque déplacement la séquence des *n* caractères est extraite, l'ensemble de ces séquences constitue l'ensemble de tous les n-grammes pouvant être générés (MILLER *et al.*, 1999).

Par exemple, les premiers 3-grammes extraits à partir de la phrase « Ceci est un exemple. » sont : *Cec*, *eci*, *ci_*, *i_e*, *es, est, st_*, …

Il y a plusieurs avantages à l'utilisation de techniques basées sur les n-grammes dont les principaux sont les suivants :

- L'analyse comparative avec d'autres techniques permet de constater que les n-grammes capturent les connaissances des mots les plus fréquents ;
- Les n-grammes opèrent indépendamment des langues, et la plupart des techniques de n-grammes n'exige pas une segmentation préalable du texte en mots, caractéristique intéressante pour le traitement de langues pour lesquelles les frontières entre mots ne sont pas fortement marquées comme le chinois ou encore les séquences ADN ;
- Les n-grammes sont tolérants aux déformations causées par l'utilisation des systèmes de reconnaissance de caractères ainsi qu'aux fautes d'orthographes (JALAM et TEYTAUD, 2000).

De manière générale, nous utiliserons la notion de *terme* pour désigner un descripteur textuel quelconque (e.g. mot, n-gramme). Nous rappelons que l'ensemble de ces termes forme l'ensemble des exogènes Γ, et que le j^{ème} terme correspondant au descripteur X^j est noté j en l'absence d'ambiguïté.

1.3.4.2 Les pondérations

La matrice des données D est constituée dans le cadre des données textuelles par le nombre des occurrences x_i^j du terme $j \in \Gamma$ apparaissant dans le texte i du corpus Ω. Cette information élémentaire peut être pondérée en fonction de divers paramètres liés à chacun des textes (e.g. le nombre de termes par texte) ou au corpus en sa totalité (e.g. le nombre de termes du corpus). La pondération permet ainsi de mieux exploiter l'information pouvant amener par exemple à l'accroissement des performances d'un système de recherche documentaire (SPARCK-JONES, 1972). La valeur ainsi pondérée, notée ϖ_i^j, est le résultat d'une application à définir sur l'occurrence x_i^j du terme j pour le texte i.

Il existe pléthore systèmes de pondération dans la littérature, mais en général ils reposent sur les deux hypothèses empiriques suivantes (AAS et EIKVIL, 1999) :

1. Plus le nombre d'occurrences d'un terme apparaît dans un texte, alors plus ce terme est important pour l'étiquette associée ;
2. Plus le nombre d'occurrences d'un terme apparaît dans le corpus, alors moins ce terme peut discriminer les documents.

Présentons les pondérations les plus usuelles avec leurs principales propriétés :

- Une première approche consiste à n'utiliser que le nombre d'occurrences des termes. Cette pondération ne peut être mise en œuvre que dans des documents de taille similaire, car sinon elle privilégierait les termes apparaissant fréquemment dans les plus longs textes :

$$\varpi_i^j = x_i^j \tag{1.3.5}$$

- Plus simplement, nous pouvons considérer que l'information importante consiste à savoir si un terme est présent ou non au sein d'un document :

$$\varpi_i^j = \begin{cases} 1 \text{ si } x_i^j > 0 \\ 0 \text{ sinon} \end{cases} \tag{1.3.6}$$

- Une autre méthode simple va consister à utiliser la fréquence relative d'un terme par rapport au nombre de termes composant un document. L'objectif ici est de permettre une comparaison entre des documents de taille différente :

$$\varpi_i^j = \frac{x_i^j}{x_\bullet^\bullet} \text{ avec } x_\bullet^\bullet = \sum_{j=1}^{p} x_i^j \tag{1.3.7}$$

- Les trois premières pondérations vérifient la première hypothèse empirique, mais nullement la seconde. A ce titre, la pondération TF×IDF corrige la fréquence du terme (*Term Frequency*) en fonction de sa fréquence au sein du corpus (*Inverse Document Frequency*). La correction se fait par multiplication du ratio des n documents du corpus par rapport au nombre de documents contenant le terme j. Le logarithme permet de lisser les résultats :

$$\varpi_i^j = x_i^j \log\left(\frac{n}{x_\bullet^j}\right) \text{ avec } x_\bullet^j = \sum_{i=1}^{n} x_i^j \tag{1.3.8}$$

- Le TF×IDF ne permet pas de prendre en compte la taille des documents. Le TFC normalise le TF×IDF en fonction de l'ensemble des termes du document, correspondant à la somme sur l des x_i^l dans (1.3.9), modulés eux-mêmes en fonction de leur proportion au sein du corpus :

$$\varpi_i^j = \frac{x_i^j \log\left(\frac{n}{x_\bullet^j}\right)}{\sqrt{\sum_{l=1}^{p}\left[x_i^l \log\left(\frac{n}{x_\bullet^l}\right)\right]^2}} \tag{1.3.9}$$

- La pondération LTC est fortement similaire au TFC, à l'exception de l'utilisation du logarithme afin d'atténuer les différences de nombre d'occurrences :

$$\varpi_i^j = \frac{\log\left(x_i^j+1\right)\log\left(\frac{n}{x_\bullet^j}\right)}{\sqrt{\sum_{l=1}^{p}\left[\log\left(x_i^l+1\right)\log\left(\frac{n}{x_\bullet^l}\right)\right]^2}} \tag{1.3.10}$$

- La pondération basée sur l'entropie est la plus sophistiquée et, d'après les expérimentations de DUMAIS (1991), est la plus performante comparée à six autres méthodes (dont le TF×IDF) : l'entropie surpasse le TF×IDF sur les cinq corpus testés. Cependant, cette méthode est de complexité importante car faisant intervenir l'ensemble des autres documents. Dès lors son intérêt lors de la nécessité de calculs rapides en est fortement réduit :

$$\varpi_i^j = \log\left(x_i^j+1\right)\left(1+\frac{1}{\log(n)}\sum_{i=1}^{n}\left[\frac{x_i^j}{x_\bullet^j}\log\left(\frac{x_i^j}{x_\bullet^j}\right)\right]\right) \tag{1.3.11}$$

avec $\frac{1}{\log(n)}\sum_{i=1}^{n}\left[\frac{x_i^j}{x_\bullet^j}\log\left(\frac{x_i^j}{x_\bullet^j}\right)\right]$ l'entropie du terme j, variant de -1 lorsque j est équidistribué sur l'ensemble des documents à 0 lorsqu'il n'apparaît que dans un document.

1.4 Sélection de variables

Dans la section précédente, nous avons décrit plusieurs méthodes permettant d'obtenir un ensemble Γ de p descripteurs permettant de représenter les données à travers D. Toutefois, un descripteur $j \in \Gamma$ n'est pas toujours pertinent voire nuisible pour l'analyse de D. Nous y voyons trois circonstances à cela :

1. Un descripteur n'est pas nécessairement utile à un problème posé. Intrinsèquement le descripteur peut ne pas contenir d'informations : le descripteur est quasi constant ou il est de mauvaise qualité car comportant des valeurs aberrantes. Le descripteur n'apporte pas d'information complémentaire par rapport aux autres descripteurs comme par exemple dans le cas de corrélations. L'utilisation de la statistique descriptive uni- ou bi-variée (e.g. moyenne, écart-type, boîte à moustache) est alors une étape à ne pas négliger car permettant à l'utilisateur de traiter rapidement ce type de descripteurs indésirables ;
2. Un descripteur peut, dans un cadre supervisé, ne pas être utile pour répondre à un problème spécifique alors qu'il pourrait être très important pour d'autres problèmes. L'évaluation des descripteurs se fait généralement de façon univariée en évaluant si l'apport informationnel de j par rapport au descripteur endogène est supérieur à un certain seuil. Des méthodes couramment utilisées sont avec l'approche univariée le Gain Ratio (QUINLAN, 1993), l'Entropie de Shannon (SHANNON et WEAVER, 1949), et avec l'approche multivariée RELIEF (KONONENKO, 1994), … Il est entendu que ces méthodes sont à adapter en fonction de la nature des descripteurs (numérique ou catégorielle) ;
3. Enfin, nous avons vu que potentiellement nous pouvions obtenir un grand nombre de descripteurs, typiquement le cas de l'image et du texte. L'objectif principal est alors d'obtenir un nombre p' de descripteurs tel que $p' \ll p$ afin que le sous-ensemble de données engendré soit traitable par des méthodes de fouilles de données. L'approche se ramène au cas précédent en sélectionnant les p' meilleurs descripteurs au sens du critère choisi. Nous étudierons des critères spécifiquement dédiés au texte en section 1.6.2.

1.5 Construction de variables

Dans l'ensemble des descripteurs, certains d'entre eux peuvent contenir de l'information pertinente pour traiter les données D mais sans être appropriés à la méthode utilisée ou au besoin de l'utilisateur. Par exemple, l'histogramme d'une image est défini par autant de variables qu'il y a de niveaux de gris. Dans certains cas, il peut être intéressant de condenser l'information en agrégeant certains niveaux contigus. Il existe principalement deux familles de méthodes pour aborder cet aspect (ZIGHED et RAKOTOMALALA, 2002) :

1. La transformation de descripteurs. L'objectif est de transformer un descripteur particulier en un autre plus approprié aux objectifs de l'analyse. Une des approches les plus usuelles est de transformer des descripteurs continus en descripteurs qualitatifs par discrétisation. Il existe nombre de ces méthodes mais le principe général consiste dans le découpage de leur domaine de valeurs en intervalles. Une autre approche de transformation consiste à centrer par rapport à la moyenne et réduire par l'écart-type les valeurs des descripteurs continues. L'intérêt de l'opération est double. Il permet de diminuer le bruit de mesure inter-individus puisque les valeurs sont considérées par rapport à l'écart à la moyenne et il confère certaines propriétés mathématiques intéressantes lors de la mise en œuvre de méthodes d'analyse de données multidimensionnelles ;
2. La construction d'agrégats. Un agrégat de descripteurs est un nouveau descripteur obtenu selon une combinaison précise des descripteurs agrégés. Par exemple, le prix au mètre-carré d'un appartement, défini par le rapport entre le prix de l'appartement et la surface totale de l'appartement,

fournit une indication assez pertinente pour comparer les appartements ou les quartiers dans les bases de données spatiales. Usuellement, les combinaisons utilisées pour l'agrégation sont les opérateurs booléens et les opérateurs mathématiques élémentaires, mais nous pouvons imaginer une multitude d'autres façons d'en obtenir. Les méthodes factorielles telles que l'analyse en composantes principales (ACP) ou l'analyse des correspondances multiples (ACM) sont largement utilisées dans ce cadre. Tout comme il n'y a pas de méthode permettant de dire qu'un descripteur est meilleur qu'un autre dans l'absolu, il n'y a pas de règles précises pour dire que tel agrégat est meilleur qu'un autre. C'est la connaissance du domaine par l'expert qui peut guider dans la définition des bons agrégats. Un agrégat peut être évalué a posteriori en respectant les mêmes procédés que pour la sélection de descripteurs « simples ».

1.6 Caractéristiques et traitements spécifiques des données textuelles

Au sein de cette section, nous allons aborder plus en détails les données textuelles. Ces données, complexes par définition, ont un fort niveau sémantique et peuvent être considérées comme un outil de représentation et de communication de la pensée. Dans un premier temps (section 1.6.1), nous allons aborder les caractéristiques linguistiques qui rendent ces données délicates à manipuler. Puis, en liaisons plus ou moins directes de ces considérations linguistiques, nous devons faire face au besoin de sélectionner un sous-ensemble pertinent de termes ou concepts (section 1.6.2).

1.6.1 Ecueils linguistiques

A la différence des données numériques, les données textuelles sont sémantiquement riches car produites, réfléchies par la pensée humaine. A la différence des langages informatiques, la langue naturelle contient nombre de règles grammaticales violées par des exceptions et surtout, comme le décrit LEFEVRE (2000), est équivoque entraînant plusieurs façons d'exprimer la même chose (redondance, synonymie) ou encore entraînant plusieurs interprétations d'un même propos (polysémie, ambiguïté).

Dans cette sous-section, nous détaillons les conséquences des principaux écueils linguistiques que nous appelons « les maux des textes » dans le cadre de la représentation vectorielle des textes. En même temps, nous rattachons ces problématiques aux divers axes de recherche s'intéressant à les traiter.

1.6.1.1 Redondance et synonymie

La redondance et la synonymie permettent d'exprimer le même concept, le même propos, à quelques nuances près, par l'utilisation de mots, d'expressions ou de phrases différentes. Lors d'une représentation vectorielle d'un texte, ces éléments sont différenciés (pratiquement ils sont dans deux colonnes

différentes), et les occurrences du concept sont éparpillées à travers toutes les formes l'exprimant. Il est alors important de pouvoir regrouper ces termes en une classe sémantique commune, par exemple regrouper les mots *jouer* et *s'amuser* dans la classe sémantique *se détendre*.

A la différence des déformations morphologiques, où l'emploi de procédures de stemming ou lemmatisation est envisageable, les relations entre les synonymes sont purement sémantiques. En outre, ce problème se complexifie lorsque des expressions paraphrasent des mots. L'approche du Traitement Automatique de la Langue Naturelle (TALN) se base sur l'utilisation de dictionnaires, appelés *thesaurus*, regroupant l'ensemble de ces relations sémantiques. L'un des plus connus est le réseau sémantique *WordNet* (MILLER *et al.*, 1990) qui regroupe d'un point de vue général les relations sémantiques, notamment la synonymie et l'hyperonymie (relation d'inclusion sémantique), de plus de 126 000 mots regroupés dans 91 000 classes conceptuelles.

Cependant, cette approche peut se révéler trop générale. Il peut alors être intéressant d'élaborer une ontologie afin de déterminer le sens des termes ou expressions employés au sein d'une organisation, d'une communauté ou d'un métier (GUARINO, 1998). Naturellement, cela engendre un coût supplémentaire quant à sa réalisation et à sa maintenance.

1.6.1.2 Ambiguïté et polysémie

Comme nous l'avons déjà dit, la langue naturelle étant univoque un mot possède souvent bien plus d'un sens et donc autant de définitions lui sont attachées. Par exemple, le mot *jouer* peut vouloir signifier *se détendre* ou bien *faire du théâtre*. En outre, l'homographie (deux mots sont dits homographes si ils s'écrivent de la même façon sans nécessairement se prononcer identiquement) est source supplémentaire d'ambiguïté. Par exemple : nous *portions* des *portions* de gâteau.

Cette ambiguïté génère du bruit et se traduit dans le cadre de la recherche d'information en une diminution de la précision (cf. Chapitre 4, section 4.2.2) des réponses retournées par le système (DE LOUPY, 2000, p. 21). Lever ces ambiguïtés semble alors nécessaire. La connaissance des catégories grammaticales des termes, obtenues par des étiqueteurs comme *Tree Tagger* (SCHMID, 1994), *FASTR* (JACQUEMIN, 2001) ou celui de BRILL (1994), permet de lever un grand nombre de ces ambiguïtés.

1.6.1.3 Asymétrie présence-absence des termes

L'utilisation de termes dans un document textuel révèle une notion / un concept que l'auteur a voulu(e) exprimer. Nous avons vu précédemment que les notions désignées par ces termes pouvaient être ambiguës mais qu'en fonction de son contexte, cette ambiguïté pouvait tout au moins être partiellement levée. Nous avons donc une relation d'implication entre un terme et son concept associé. Néanmoins, nous avons également constaté qu'il y avait plusieurs moyens d'exprimer les mêmes notions. Dès lors, l'absence d'un terme n'implique pas nécessairement que la notion qui lui serait liée se trouve être absente du document.

Cette remarque triviale a pour conséquence la vigilance quant à l'utilisation de règles d'apprentissage s'appuyant sur l'exclusion d'un terme particulier.

1.6.1.4 Opérateurs de négation

Ces termes particuliers (e.g. *ne*, *non*) de la langue naturelle permettent de modifier la signification des termes associés à cette négation. Cette modification est locale dans le sens où elle affecte seulement les termes qui lui sont proches. C'est pour cela qu'il est inintéressant d'utiliser ces termes tels quels pour la représentation vectorielle des textes. En outre, la notion affectée par les opérateurs de négation, elle, reste inchangée. Par exemple les deux phrases suivantes *il fait beau aujourd'hui* et *il ne fait pas beau aujourd'hui* traitent toutes deux du *temps*, et le terme *beau*, avec ou sans négation, est un terme décrivant cette notion de *temps*. Naturellement, elles ont une signification opposée, mais sont toutes deux rattachées à la thématique du *temps*.

La prise en compte de ces opérateurs n'est pas toujours une nécessité, car dépendante du niveau de détails requis par l'objectif du traitement textuel exécuté. Dans le cadre d'une recherche d'information spécifique ou recherche documentaire[1], l'objectif est pour l'utilisateur de rechercher de l'information en adéquation avec la signification de sa requête exprimée en langue naturelle. Par contre, dans le cadre d'une catégorisation de documents textuels en plusieurs thématiques, les éléments de négation ne vont guère jouer un rôle essentiel puisque l'on cherche à distinguer les thématiques les unes des autres.

[1] Ce point sera détaillé au Chapitre 4

1.6.1.5 L'apport du traitement automatique de la langue naturelle

Face aux problèmes liés à la langue naturelle, des traitements linguistiques peuvent être mise en œuvre. Un premier inconvénient réside en la nécessité de disposer de ressources (dictionnaires, ontologies, …) spécifiques à la langue mais également à la problématique. Outre ces considérations, nous pouvons nous interroger sur ce que ces opérations bien souvent coûteuses en temps apportent en termes de performances, de précision.

Dans le cadre de la recherche documentaire, DE LOUPY (2000) a montré que l'utilisation de la désambiguïsation sémantique et la gestion de la synonymie et de la polysémie peuvent apporter des résultats satisfaisants sans toutefois être suffisamment pertinentes pour conclure à la nécessité d'utiliser systématiquement ces approches. De même, dans le cadre de la catégorisation de documents, SCOTT et MATWIN (1999) ont montré que l'utilisation de connaissances linguistiques ne majorait pas les performances des classifieurs.

1.6.2 Sélection de variables textuelles

Un des principaux enjeux pour le traitement des données textuelles est de définir quels sont les termes les plus appropriés pour l'apprentissage, puis de sélectionner parmi ces termes ceux qui sont les plus significatifs. Par exemple dans un contexte supervisé, il s'agit de déterminer l'ensemble de termes qui assureraient les meilleures performances au modèle de prédiction construit.

Les méthodes de sélection de variables textuelles s'inscrivent dans un cadre particulier. En effet, le nombre initial de termes, produit par le pré-traitement du corpus, est potentiellement très élevé (p variables, avec très souvent $p > 10\,000$). Dans ce contexte, il est nécessaire que la complexité des méthodes soit linéaire en nombre de termes afin d'avoir des temps d'exploitation raisonnables. L'information principale est basée sur la fréquence d'apparition d'un terme au sein d'un corpus, c'est pourquoi les méthodes de sélection orientées données textuelles sont généralement basées sur les occurrences des termes. Cependant, comme nous le verrons plus loin en 1.6.2.1, la quantité de termes à sélectionner reste un problème ouvert auquel s'ajoute l'effet « palier » en raison d'un même score pour un sous-ensemble de termes. Nous aborderons ensuite en 1.6.2.2 une approche non supervisée permettant d'écarter rapidement les termes potentiellement non pertinents ou non utilisables.

Les méthodes couramment utilisées pour la sélection de variables en fouille de données, performantes certes, mais de complexité élevée, sont peu appropriées (LIU et MOTODA, 1998). De fait, les méthodes

de sélection de termes les plus souvent mises en œuvre en catégorisation de textes sont essentiellement univariées afin de rester de complexité $O(p)$. Ces méthodes, qui semblent donner satisfaction dans certains contextes, notamment lors de la recherche de présence ou absence d'une thématique, présentent néanmoins deux défauts. Le premier est de ne pas tenir compte des interactions entre les termes, le rôle de chaque terme étant évalué indépendamment des autres. Le second se produit dans le cas où l'étiquette à prédire peut prendre plusieurs étiquettes ; ces méthodes ne permettent pas de déterminer, lorsqu'un terme est significatif, à quelle association « catégorie » - « terme » est due sa sélection, et par là, à la reconnaissance de quelle catégorie il contribue le plus.

Nous présentons d'abord en 1.6.2.3 une méthode de sélection bien connue et largement utilisée : la sélection selon le $\chi^2_{\text{univarié}}$, suivie en 1.6.2.4 d'une proposition multivariée de sélection de termes, ayant donné lieu à une publication (CLECH *et al.*, 2003a). Puis, nous présentons en 1.6.2.5 une expérimentation qui permet d'évaluer la pertinence de notre approche par rapport à la méthode du $\chi^2_{\text{univarié}}$ faisant référence. Enfin, nous conclurons en 1.6.2.6 sur ce thème en mettant en perspective les améliorations possibles.

1.6.2.1 Quantité de termes et effet « palier »

De part la quantité importante de descripteurs potentiels, la sélection de variables est une étape particulièrement importante pour le traitement des données textuelles. Comme dans le cadre général, l'objectif est de déterminer les variables les plus informatives au sens d'un critère fixé. Cependant, peut-être encore plus ici qu'ailleurs, le choix du nombre de descripteurs à sélectionner est difficile. En effet, pour l'heure il n'existe pas de tests statistiques satisfaisants permettant d'effectuer ce choix, tant cette quantité est liée à la problématique abordée. Dès lors, pour faire face à la problématique d'espace creux (un très faible nombre d'individus pour une très forte quantité de variables, e.g. 1‰) comme définie dans (BELLMAN, 1961), une approche pragmatique consiste à évaluer la proportion de termes souhaitée pour un individu. A titre d'exemple, dans un contexte de recherche d'information FUHR et BUCKLEY (1991) proposent d'utiliser 50 à 100 fois plus de textes que de termes. Enfin, l'approche empirique, comme effectuée par (YANG et PEDERSEN, 1997), consiste à évaluer les performances obtenues avec plusieurs quantités de termes sélectionnés afin de juger l'effectif optimal face à un problème. Même si en pratique la première approche apporte une réponse relativement satisfaisante, l'inconvénient majeur est de ne pas trouver le nombre optimal de descripteurs. L'inconvénient princi-

pal de la seconde approche est le coût en calcul et en temps pour une variation relativement faible des performances.

Par ailleurs, l'information utilisée pour la sélection des données textuelles est fonction soit des fréquences des termes au sein d'un texte, soit de leur apparition ou absence. Or, il est courant de rencontrer au sein d'un corpus des termes ayant la même fréquence d'apparition impliquant alors l'obtention d'un même score par les méthodes de sélection de variables textuelles, constituant ainsi un palier (Figure 1-7). Dès lors, une sélection par nombre de termes doit s'ajuster afin de considérer un palier dans sa totalité. Pour nous rapprocher le plus du nombre d'éléments souhaités, nous contraignons l'ajustement à ajouter ou supprimer le moins d'éléments possibles pour arriver respectivement à la fin (k_1 sur la Figure) ou au début (k_0 sur la Figure) du palier.

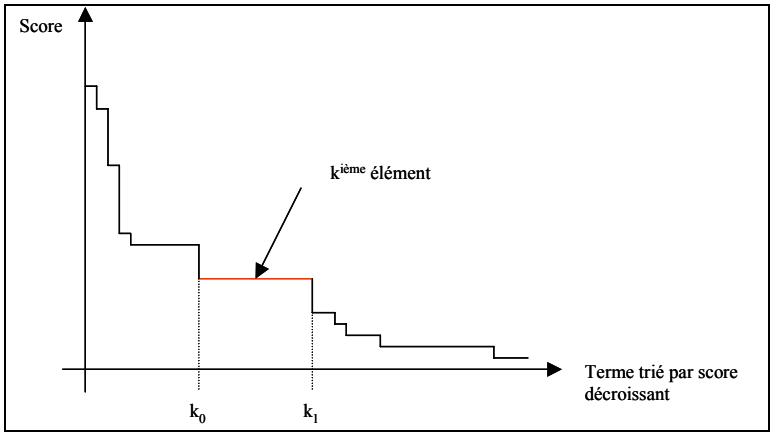

Figure 1-7 Représentation de l'effet « palier »

1.6.2.2 Sélection selon les fréquences

Une première méthode de sélection, pouvant être considérée comme une méthode de pré-traitement agressive, consiste à supprimer les mots trop fréquents et trop rares afin de conserver les mots importants (VAN RIJSBERGEN, 1979). En effet, lorsque les termes sont trop fréquents, ils apparaissent dans la quasi totalité des textes d'un corpus, et ont donc un pouvoir discriminant limité. Généralement, ces termes sont des mots grammaticaux, mots vides ou mot outils (e.g. l'ensemble des déterminants). Cependant, même si la plupart du temps ils ne sont pas à retenir, ces derniers peuvent être très précieux pour lever certaines ambiguïtés linguistiques, ou encore pour étudier la stylistique d'un auteur. Pour écarter les mots trop fréquents, nous fixons un seuil maximal de fréquence permettant de ne pas sélec-

tionner les termes présents dans une très forte proportion de textes (e.g. que le terme n'apparaisse pas dans 90 textes d'un corpus composé de 100 textes).

Lorsque les mots sont trop rares, même si sémantiquement ils peuvent être très riches, leur pouvoir discriminant est également limité. Par exemple, les hapax sont des mots qui n'apparaissent qu'une seule fois dans le corpus. Les inclure en tant qu'index n'apporte que très peu d'informations. Ainsi, nous définissons d'une part un seuil d'occurrences minimal pour écarter les termes trop rares du corpus (e.g. au moins 5 occurrences dans tout le corpus). D'autre part, nous fixons un seuil minimal de fréquence permettant d'évacuer les termes seulement présents dans une très faible quantité de textes du corpus (e.g. que le terme apparaisse au moins dans 5 textes d'un corpus composé de 100 textes).

1.6.2.3 Sélection par le χ^2 univarié

La statistique du $\chi^2_{\text{univarié}}$ mesure l'écart à l'indépendance entre un terme $X^j \in \Gamma$ et une étiquette $e_k \in \mathcal{E}$. Cette mesure s'élabore à partir d'une table de contingence (Tableau 1-1) indiquant le nombre de documents appartenant à l'étiquette e_k où le terme X^j est soit présent (soit a documents dans le Tableau), soit absent (soit c documents dans le Tableau). Nous faisons de même avec les documents n'appartenant pas à e_k, nous obtenons alors les quantités b et d. Dès lors, cette table de contingence doit être construite pour chacun des termes candidats, et sa marge totale est par construction égale au nombre de documents du corpus. Cette étape ne nécessite qu'une seule analyse du corpus, rendant ainsi la complexité linéaire au nombre de termes.

	e_k	$\overline{e_k}$	
X^j	a	b	$a+b$
$\overline{X^j}$	c	d	$c+d$
	$a+c$	$b+d$	$n=a+b+c+d$

Tableau 1-1 Table de contingence selon le nombre de documents

A partir de cette table, la statistique du χ^2 peut se mettre sous la forme suivante :

$$\chi^2_{\text{univarié}}\left(x^j, e_k\right) = \frac{n\left(ad-cb\right)^2}{\left(a+c\right)\left(b+d\right)\left(a+b\right)\left(c+d\right)} \tag{1.6.1}$$

Lorsqu'un terme X^j apparaît exclusivement dans les documents portant l'étiquette e_k, alors les quantités b et c sont nulles, entraînant ainsi une valeur de $\chi^2_{\text{univarié}}\left(X^j, e_k\right) = n$. A l'autre extrême, lors de l'indépendance entre le terme X^j et l'étiquette e_k, se traduisant par la même fréquence relative de X^j dans l'ensemble des documents portant l'étiquette e_k que dans l'ensemble des documents portant une étiquette différente, alors les quantités ad et bc sont équivalentes entraînant ainsi une valeur de $\chi^2_{\text{univarié}}\left(X^j, e_k\right) = 0$. Entre ces deux extrêmes, plus la valeur du $\chi^2_{\text{univarié}}\left(X^j, e_k\right)$ est grande, plus X^j et e_k sont liés.

Afin de mesurer globalement l'information apportée par le terme X^j, deux associations des scores obtenus pour chaque étiquette sont habituellement utilisées : la moyenne, notée $\chi^2_{\text{moyenne}}\left(X^j\right)$ (1.6.2), ou le maximum, noté $\chi^2_{\max}\left(X^j\right)$ (1.6.3) (YANG et PEDERSEN, 1997).

$$\chi^2_{\text{moyenne}}\left(X^j\right) = \sum_{k\in[1,m]} \frac{\left|e_k\right|}{n} \chi^2_{\text{univarié}}\left(X^j, e_k\right) \tag{1.6.2}$$

Avec $\left|e_k\right|$ représentant le nombre de documents ayant l'étiquette *k*.

$$\chi^2_{\max}\left(X^j\right) = \max_{k\in[1,m]} \left\{\chi^2_{\text{univarié}}\left(X^j, e_k\right)\right\} \tag{1.6.3}$$

En pratique, comme illustré dans (YANG et PEDERSEN, 1997), le $\chi^2_{\max}\left(X^j\right)$ est le plus utilisé des deux, apportant généralement de meilleurs résultats dans le cadre de la catégorisation ayant un nombre d'étiquettes strictement supérieur à 2. L'une des raisons à cela est que le $\chi^2_{\max}\left(X^j\right)$ requiert seulement que X^j soit fortement lié à une étiquette particulière, alors que pour le $\chi^2_{\text{moyenne}}\left(X^j\right)$ cette liaison forte risque d'être atténuée de part la moyenne sur l'ensemble des classes.

1.6.2.4 Sélection par le χ^2 multivarié

Le χ^2 multivarié, noté $\chi^2_{\text{multivarié}}$, est une méthode supervisée permettant la sélection de termes en prenant en compte non seulement leurs fréquences dans chaque classe mais aussi l'interaction des termes entre eux et les interactions entre les termes et les classes. Le principe consiste à extraire les ξ

meilleurs termes caractérisant le mieux une classe par rapport aux autres, ceci pour chaque classe, ξ étant fixé par l'utilisateur.

Pour ce faire, le tableau croisé global (termes – étiquettes) du nombre total d'occurrences des termes, de dimension $p \times m$, est calculé (Tableau 1-2). La somme totale des occurrences est notée *N*. Les valeurs N_{jk} des cellules $\left(X^j, e_k\right)$ représentent le nombre de fois où le terme X^j est présent dans les documents étiquetés e_k. Puis, les contributions de ces cellules $\left(X^j, e_k\right)$ au χ^2 associé à ce tableau sont calculées comme indiqué dans l'équation (1.6.4), puis triées par ordre décroissant pour chacune des classes. L'évaluation du signe dans l'équation (1.6.4) permet de déterminer le sens de la contribution du terme à la classe : une contribution positive indique que c'est la présence du terme qui y participe tandis qu'une contribution négative révèle que c'est son absence qui y participe.

Les principales caractéristiques de cette méthode sont les suivantes :

- Elle est supervisée car elle s'appuie sur l'information apportée par l'ensemble étiquette $\mathcal{E}$;
- Elle est multivariée car elle évalue globalement le rôle d'un terme par rapport aux autres ;
- Elle tient compte de l'interaction termes-classes car elle permet de choisir, pour chaque catégorie, les termes qui contribuent le plus à leur discrimination ;
- Malgré sa sophistication, elle reste de complexité linéaire en nombre de termes.

	e_1		e_k		e_m	
X^1	N_{11}	…	N_{1k}	…	N_{1m}	$N_{1\bullet}$
⋮	⋮		⋮		⋮	
X^j	N_{j1}	…	N_{jk}	…	N_{jm}	$N_{j\bullet}$
⋮	⋮		⋮		⋮	
X^n	N_{n1}	…	N_{nk}	…	N_{nm}	$N_{n\bullet}$
	$N_{\bullet 1}$		$N_{\bullet k}$		$N_{\bullet m}$	$N = N_{\bullet\bullet}$

Tableau 1-2 Tableau croisé global du nombre total d'occurrences

$$C_{jk}^{\chi^2} = N \frac{\left(f_{jk} - f_{j\bullet} f_{\bullet k}\right)^2}{f_{j\bullet} f_{\bullet k}} \times \text{signe}\left(f_{jk} - f_{j\bullet} f_{\bullet k}\right) \tag{1.6.4}$$

Avec $f_{jk} = \frac{N_{jk}}{N}$, représentant les fréquences relatives des occurrences

1.6.2.5 Expérimentation

Afin d'illustrer la pertinence de notre approche, nous comparons nos résultats, en validation croisée, du modèle de catégorisation de dépêches du journal Le Monde de 1994 avec ceux obtenus en utilisant une méthode univariée de référence : le χ^2_{max}. Nous rappelons que la nature de l'information utilisée est très différente dans ce cas. En effet, dans le cadre multivarié l'information utilisée est le nombre d'occurrences des termes (la marge totale équivaut à la somme du nombre d'occurrences des termes sur tout le corpus), tandis que dans le cadre univarié, l'information est le nombre de documents où un terme apparaît (la marge totale est le nombre de documents).

Le corpus Le Monde est constitué de 419 dépêches, divisées en 10 catégories. La quantité de dépêches par catégorie est très hétérogène allant de 19 pour la plus petite à 95 pour la plus importante. Les catégories sont des sujets définis par des experts regroupant les dépêches s'y référant (par exemple Téléphone Portable, Conflit en Palestine). En outre, une dépêche n'est assignée qu'à une et une seule catégorie.

La méthode d'apprentissage utilisée est usuelle dans le cadre de la catégorisation : le 3 Plus Proches Voisins (3-PVV) ; cette méthode sera décrite dans le chapitre suivant (cf. 2.2). Ce choix est également motivé par la sensibilité de cette méthode à la qualité de l'espace de représentation, c'est-à-dire les termes sélectionnés pour l'apprentissage. Nous avons paramétré cette méthode en utilisant les votes pondérés par l'inverse de leur distance et la métrique cosinus (AMINI, 2001). Les valeurs des termes sélectionnés sont quant à eux pondérés par le TF×IDF, là encore largement utilisé pour un problème de catégorisation.

Notre chaîne de traitements consiste donc en une étape de sélection de termes (multivariée ou univariée) dont les valeurs des termes sélectionnés sont pondérées par le TF×IDF, puis en une d'apprentissage réalisée par le 3 PPV. En nous appuyant sur la validation croisée (appliquée à l'ensemble de la chaîne de traitement), nous mesurons alors plusieurs indicateurs d'évaluation de la catégorisation de textes (Figure 1-8) : le taux de succès et son écart-type, le rappel macro-moyen et la précision macro-moyenne (moyenne des rappels, respectivement des précisions, sur l'ensemble des catégories) ; ces indicateurs seront décrits dans le Chapitre 2 (cf. 2.4).

Enfin, dans le but de pouvoir comparer précisément les performances des deux méthodes, nous n'avons pas activé la prise en compte de l'effet « palier » qui aurait pu faire fluctuer la quantité de termes sélectionnés, et par là même biaiser les performances des classifieurs.

Les résultats de la Figure 1-8 montrent que pour chaque type de termes utilisés (mot, 3-, 4- et 5-grammes), la méthode du $\chi^2_{\text{multivarié}}$ amène à une meilleure qualité de prédiction, quel que soit l'indicateur de qualité utilisé pour évaluer l'apprentissage. En outre, nous remarquons la constance des résultats du $\chi^2_{\text{multivarié}}$ pour chacun des indicateurs indépendamment du type de termes.

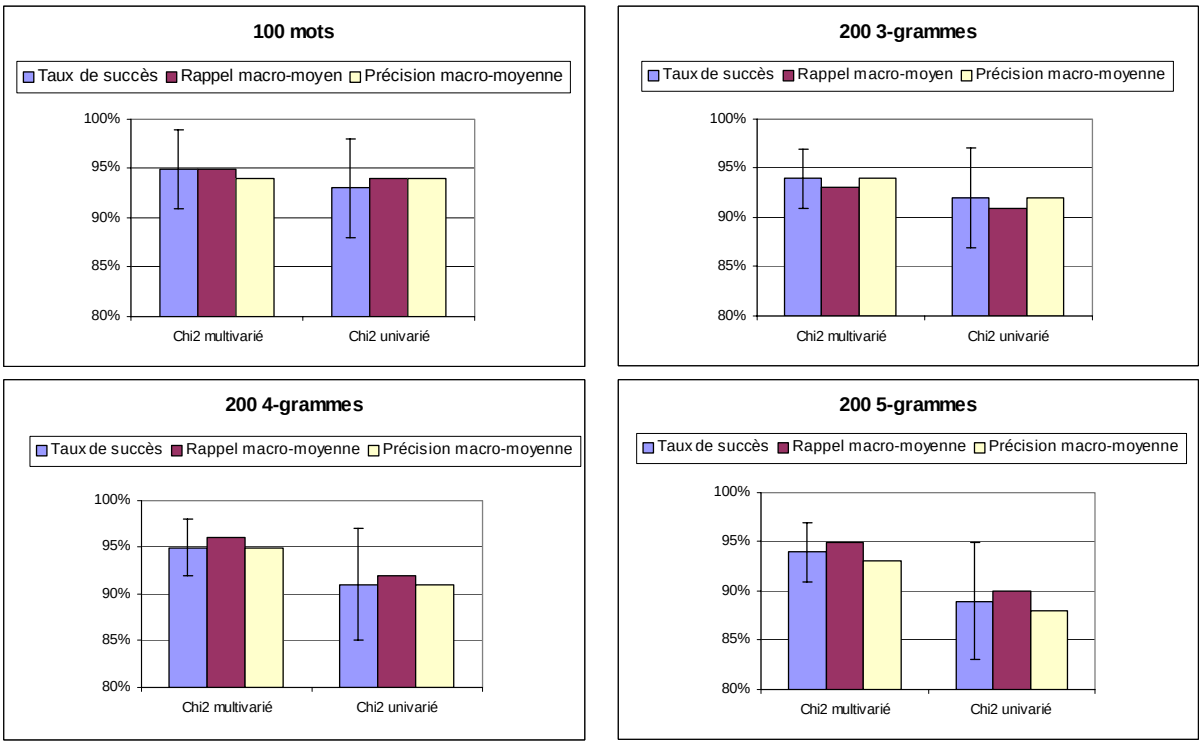

Figure 1-8 Qualité du modèle en 10 validations croisées
selon la nature des termes utilisés et la méthode de sélection

1.6.2.6 Bilan de la sélection de termes

Nous avons proposé une méthode originale pour la sélection de termes lors de la catégorisation de textes. Cette méthode, de complexité linéaire, s'appuie sur la contribution du χ^2 bien connu en statistique. Elle se démarque des méthodes univariées usuelles par la nature de l'information utilisée : elle est fondée sur la fréquence des termes et non leur présence/absence. De plus, elle tient compte des interactions entre les termes et des interactions termes/catégories. Nos premières évaluations indiquent que l'approche est efficace, nous devons néanmoins approfondir nos expérimentations pour bien discerner les avantages et inconvénients des différentes caractéristiques de l'algorithme. Cela nous permettra par la suite de mieux définir, et donc de mieux reconnaître, les situations où l'approche

s'avérera la plus intéressante. Enfin, si à l'heure actuelle l'algorithme repose sur un paramètre *ad hoc*, i.e. le nombre de termes à sélectionner pour chaque catégorie (devant être fixé par l'utilisateur), l'étude d'un test statistique de la significativité des contributions nous semble une piste intéressante afin de déterminer automatiquement la quantité de termes les plus pertinents pour la discrimination.

1.7 Conclusion

Nous avons introduit dans ce chapitre un cadre général à l'Extraction de Connaissances à partir de Données Complexes. Nous avons, à travers de multiples exemples, décrit la démarche d'extraction de caractéristiques au sein de documents complexes de diverse nature dans un contexte de fouille : nous avons exprimé les motivations, les choix et les méthodes pour y parvenir.

Nous nous sommes ensuite intéressés plus particulièrement aux documents textuels. Nous avons décrit la difficulté de manipuler ces documents en raison de leur riche contenu sémantique. Nous avons montré cependant que plusieurs approches permettent de contourner ces écueils. Puis, nous avons rappelé le problème de la forte dimensionalité des descripteurs textuels et exposé des méthodes usuelles de sélection. Enfin, nous avons proposé une nouvelle méthode de sélection de termes prenant en compte une information plus riche et apportant des résultats systématiquement meilleurs à la méthode comparée dans les diverses conditions de tests.